Crystal Structure Prediction: Are we there yet?
Aurora J. Cruz-Cabeza
The University of Manchester, Sackville Street, Manchester, M13 9PL (UK)
email: aurora.cruzcabeza@manchester.ac.uk

Predicting the way a molecule will crystallise under a given set of conditions is a challenging task. Sketch out a molecule into a computer and run complex algorithms in the hope that the computational result matches that of the experiment. Being able to predict crystal structures computationally in an accurate and time-efficiently manner could have enormous implications in industries that develop organic crystalline materials (e.g. pharmaceuticals). Crystal Structure Prediction (CSP) is, therefore, an ultimate dream in the context of pharmaceutical material sciences, but are we there yet?

The development of CSP methods has been partly catalysed by the blind test competitions organised at the Cambridge Crystallographic Data Centre. Experimental outcomes of crystallisations for a set of targets are held in confidence whilst molecular sketches are given to participants who are then allowed to run their computations over the course of nine months and submit their computer-generated crystal structures. The blind test was first run in 1999 and since, it has been regularly organised every three years. The results of the sixth blind test of CSP methods are being published in the present issue of Acta Crystallographica. So, what have we learnt?

The CSP problem can be broken into two parts: i) a sampling problem and ii) a ranking problem. The sampling problem relates to the number of ways a given molecule can be packed in a three-dimensional space in the form of an ordered crystal. The ranking problem relates to the identification of the structure or structures, which will be observed experimentally from all those possible packings under a given set of conditions. Over the years, blind test targets have increased in complexity; but, what does complexity mean in this context? Sampling becomes more difficult as we increase the flexibility of the target molecule, as we increase the number of symmetrically independent molecules in the crystal and as we increase the number of components. A complex sampling, however, does not actually mean a complex ranking too. In fact, quantifying complexity at the ranking stage, a priori, is virtually impossible.

Figure 1. Targets in the sixth blind for crystal structure prediction.

In the sixth blind test, five different targets were attempted (Figure 1): a small rigid molecule (XXII), two considerably large and flexible molecules (XXIII and XXVI), a two-component cocrystal (XXV) and a hydrate of a salt (XXIV). In addition, target
XXIII had the difficulty of crystallising in five different packings (polymorphs), two of which had two symmetrically independent molecules.

As in all previous blind tests, methods accuracy and group experience played an important role in the success of the groups. The know-how is an important element as well as the quality of the sampling and ranking methods. Participants trying to implement newer methods often only attempted the simplest target XXII whilst only groups with more history and experience attempted all targets.

Consolidated sampling algorithms were able to generate the observed packing in all targets except for the salt hydrate, which was only generated by one group, and the structure of one of the polymorphs of target XXIII, which was not generated by any method. The sampling complexity of this blind test has been enormous and tremendous progress has been achieved, however, it has been highlighted (once again) that we still have some way to go. Hydrates of salts are rather common in pharmaceuticals as well as complex polymorphs. Given a molecule, the sampling of all possible solid forms is virtually infinite once we start accounting for polymorphs, cocrystals, salts, hydrates of salts, solvates of cocrystals... Whilst consolidated algorithms are now able to sample the packing of the some of these forms (many of which are incredibly complex), we are still far from being able to do efficient computational screenings of the entire solid form universe of a given compound. We are at the point, however, at which these methods can complement and add value to experimental screenings.

With respect to rankings, most ranking methods are solely based on energy (lower energy structures are therefore more likely to be observed experimentally). It was well established in previous blind tests that classic force field models perform poorly at the ranking stage, so most groups have now adopted more sophisticated energy models including some based on density functional theory. Most targets would have been correctly ranked by various of the more sophisticated ranking methods, some of which have been intensively developed in the past years. This is a very impressive outcome, which is consistent with the findings in the fourth and fifth blind test. The most stable polymorphs for target XXIII (A & D), albeit only identified under unusual conditions, were not correctly ranked by any method. In that regard, whilst there has been a tremendous advancement in ranking methods, further studies with complex polymorphic systems are a must for future editions of the blind tests and for future benchmarking studies. As discussed above, it is not yet clear when a particular target is going to present an especially difficult ranking stage. Why are most targets perfectly ranked with advanced energy methods but a few targets still misbehave?

Finally, we are still far from relating experimental conditions of crystallisation to the obtained crystal structures. In fact, none of the blind test participants made use of the crystallisation conditions for their predictions. In an ideal scenario, we would like to be able to predict structures computationally and then derive an experimental procedure to produce them in the laboratory. For this to occur, however, there is a considerable amount of fundamental work still needed to advance our understanding of nucleation and our ability to simulate it computationally. Whilst we are coming close to predicting the stable structure of a given compound, we still cannot answer why some compounds are polymorphic or which of the predicted forms can be realised experimentally and how. As the methods have been advancing, the actual purpose of the blind test has been evolving too from prediction of structure to prediction of solid-form landscapes and polymorphism.

The sixth blind test, more than ever before, represents an outstanding community effort with twenty-five groups, ninety-two authors and fifty-two institutions taking...
part. Beside that, extensive experimental form screenings were organised for two of the targets. Wow! This is a must read article!